DOCUMENT RESUME

ED 263 123 TM 850 558

AUTHOR Huberty, Carl J.
TITLE On Statistical Testing.
PUB DATE Apr 85
NOTE 27p.; Paper presented at the Annual Meeting of the American Educational Research Association (69th, Chicago, IL., March 31-April 4, 1985).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Goodness of Fit; Higher Education; *Hypothesis Testing; Mathematics Materials; Research Design; *Statistical Significance; Statistical Studies; Statistics; Textbook Evaluation
IDENTIFIERS *Fisher Exact Probability Test; *Neyman Pearson Hypothesis Testing; P Values

ABSTRACT

An approach to statistical testing, which combines Neyman-Pearson hypothesis testing and Fisher significance testing, is recommended. The use of P-values in this approach is discussed in some detail. The author also discusses some problems which are often found in introductory statistics textbooks. The problems involve the definitions of hypothesis, region of rejection, contingency tables, and tails of a test. A reference list is appended. (Author/GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED263123

On Statistical Testing

Carl J Huberty

University of Georgia

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Huberty, Carl

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

Paper presented at the annual meeting of the American Educational Research Association, Chicago, April 1985.

TM 850558

Abstract

A statistical testing approach that is a hybrid of (Neyman-Pearson) Hypothesis Testing and (Fisher) Significance Testing is advanced. An aspect of the approach, the use of P-values, is discussed in some detail. Questionable textbook use of some statistical testing terminology is reviewed.

On Statistical Testing

Statistical Testing Approaches..... 1

Neyman-Pearson Hypothesis Testing..... 2
Fisher Significance Testing..... 4
On Alpha-Values..... 5
On P-Values..... 5

A Hybrid Approach..... 9

On Some Statistical Testing Terminology..... 14

Hypothesis..... 14
Region of Rejection..... 15
Tails of a Test..... 16
Contingency Table..... 16

Discussion..... 17

References..... 20

On Statistical Testing

The motivation for a discussion of statistical testing arose from my teaching of this topic to graduate students in introductory statistical methods classes. The motivation peaked recently when using first a textbook by Guilford and Fruchter (1978) and then a textbook by Mosteller, Fienberg, and Rourke (1983). Both sets of authors introduce statistical testing a la Fisher; the former pair follows their introduction with a discussion of the Neyman-Pearson approach, while the latter trio continues throughout essentially with the Fisher approach. Many other textbooks introduce statistical testing a la Neyman-Pearson, and then slip in the Fisher approach in a somewhat nonchalant manner.

One purpose of this paper is to briefly review the two statistical testing approaches, and then to suggest that practicing researchers should employ a combination of the two approaches in designing a study and analyzing the data. In the process of reviewing textbooks for potential use for an introductory statistical methods course, questionable use of some statistical testing terminology has been found. The second purpose is to discuss concerns with expressed meanings and use of four such terms.

Statistical Testing Approaches

The expression "statistical testing" is used herein when referring to either Hypothesis Testing or Significance Testing. These two specific terms (and/or notions underlying them) are often used without distinction in general references such as handbooks and encyclopedias, in many introductory statistical

methods textbooks, in some writings about data analysis (e.g., Carver, 1978; Meehl, 1978; Sawyer & Peter, 1983), and in some writings by "pure" statisticians (e.g., Neyman, 1976; Snedecor & Cochran, 1967). These two terms (and the notions underlying them) are, however, considered to be different by some writers: in statistics by Kempthorne (1972, 1976), in psychology by Lawlis and Peek (1979), and in philosophy by Spielman (1974), to mention a few expositions. (It is interesting, perhaps, that Hodges and Lehmann (1970, p. 317) apparently consider Significance Testing to be a "short-cut device" for Hypothesis Testing.) When the two approaches or processes are viewed as being different, the "founders" of Hypothesis Testing are generally considered to be J. Neyman and E. S. Pearson, and the founder of Significance Testing to be R. A. Fisher. (Tests of significance were, however, used long before their advocation by Fisher--see Carlson, 1976.)

No attempt will be made here to present a complete, technical, philosophical discussion of the two approaches. (One might refer to a textbook on statistics (e.g., Cox & Hinkley, 1974; Kempthorne & Folks, 1971), to the monographs by Hacking (1965) and Seidenfeld (1979), or to the articles by Carlson (1976), Kempthorne (1976), and Spielman (1974, 1978).) Rather, a description of each of the approaches will be given in a "step-wise" form. It is assumed that the reader is familiar with basic terms and notation commonly used in discussing statistical testing.

Neyman-Pearson Hypothesis Testing. Consider a research design that is a one-way layout with four groups of 20 experimental units each. Here, a one-way analysis of variance is

to be employed. The general procedural steps for Hypothesis Testing are given below, along with exemplary steps:

		Example
NP1.	State statistical hypotheses	HO: 4 pop. means equal H1: not HO
NP2.	Select appropriate test statistic	MSH/MSE
NP3.	Specify referent (i.e., null) probability distribution (assuming necessary data conditions are met)	F(3,76)
NP4.	Specify probability of a Type I error (alpha) one is willing to live with, and determine the region of rejection (R)	(alpha = .08) R: F > 2.328
NP5.	Collect data and calculate value of test statistic	F = 4.27
NP6.	Decide to reject HO (in favor of H1) or not depending on whether or not observed value of test statistic is in R	Reject HO

Although not explicitly considered in an Hypothesis Testing "step," the region of rejection (R) is determined so that the probability of getting a test statistic value in R, under H1, is as large as possible. More formally, R is determined so that P(F in R | HO) = alpha (the equality implies that HO is a

"simple" hypothesis), and that P(F in R | H1) is maximized; the latter probabilities are called power values.

Fisher Significance Testing. Considering the same research design as above, the general procedural steps for Significance Testing are given below, along with exemplary steps:

		Example
F1.	State null hypothesis	HO: 4 pop. means equal
F2.	Select appropriate test statistic	MSH/MSE
F3.	Specify referent (i.e., null) probability distribution (assuming necessary data conditions are met)	F(3,76)
F4.	Collect data and calculate value of test statistic	F = 4.27
F5.	Determine estimated probability (P) of obtaining such a test statistic value or one more extreme under HO	P = P(F(3,76) > 4.27 | HO) = .008
F6.	Conclude data do not support HO if P is "small"; otherwise HO is supported	HO not supported

Some notions that are typically not considered in Significance Testing include alternative hypothesis, region of rejection, Type I and Type II errors, and power (see Savage, 1976, pp. 470-474).

On Alpha-Values. The probability of a Type I error (alpha) in step NP4 with which one is willing to live is both researcher- and situation-dependent. True, some textbooks lead the reader to believe that the 5% level is sacrosanct. (See Cowles and Davis (1982) for an historical perspective.) It is stated in one textbook: "The .05 level is...commonly used, but ordinarily no greater risk of Type I error should be permitted." It is hoped that very few researchers believe that any alpha is sacred. The .05 level might be considered a "fence-rider's alpha," the level used when one doesn't know whether to be liberal or conservative. If one is involved in exploratory research, breaking new ground so-to-speak, perhaps an alpha of .10 or .15 or higher would be acceptable. Whereas, in research on problems and variables heavily studied in the past, perhaps a lower alpha, say, .01, would be preferred. When in a given study, multiple statistical tests are conducted, an alpha (over all tests) in the range of .10 to .20 seems to be reasonable.

One could also consider which of the two types of errors, I or II, is more consequential, and allow that consideration to dictate a "low" or "high" alpha. Still further, a "benefit-cost" approach may be considered (Nagel & Neef, 1977). Morrison and Henkel (1970, pp. 155-171) provide further reading on the choice of alpha.

On P-Values. It should be noted that the alpha in step NP4 is a fixed value, while the "P-value" in step F5 is a random variable. Their very definitions indicate that they should not be considered synonymously, while, in fact, they often are. It is for this reason, it is hypothesized, that the use of P-values in

statistical testing has been somewhat controversial. In some textbooks for researchers in the behavioral sciences (e.g., Hays, 1981; Kirk, 1984), the use and reporting of P-values is not suggested. The lack of discussion in textbooks written for behavioral science researchers is somewhat puzzling in light of the common practice of reporting P-values (the lower-case p is often used) in journal articles and in light of attention paid to them in publication manuals (e.g., APA, 1983). Are such authors trying to tell us something? Should we listen?

Descriptors used for a P-value include "significance probability" (Bhattacharyya & Johnson, 1977, p. 175; Hodges & Lehmann, 1970, p. 317), "descriptive significance level" (Mosteller et al., 1983, p. 238), "achieved level of significance" (Anderson & Sclove, 1978, p. 414), and simply "probability level" (APA, 1983).

Discussion of the use and interpretation of P-values is quite aptly given by Gibbons and Pratt (1975; augmented with letters by Brewer, 1975, and Trawinski, 1975) and by Kempthorne (1972, p. 187). P-values are generally considered as measures of implausibility of H0, as a means of quantifying the weight of evidence against H0, as a "measure of the degree of surprise which the experiment should cause a believer" of H0 (Hodges & Lehmann, 1970, p. 317), or as a measure of the distance between a given data set and the population of data sets of which the given data set is a sample. When a P-value is used in conjunction with Hypothesis Testing, it is generally considered the smallest significance level at which the null hypothesis is rejected (see, e.g., Glass & Hopkins, 1984, p. 205).

R. A. Fisher did not consider, perhaps, a significance test with more than one "tail" (see a later section in the present paper). This is not to say that Fisher considered only directional alternative hypotheses--it is not clear that he considered the notion of an alternative hypothesis at all. It is a moot point whether P-values are appropriate in the case of nondirectional alternatives. It seems reasonable, when the (null) referent distribution is symmetrical, to report a P-value for a non-directional alternative that is twice the P-value for a directional alternative. Gibbons and Pratt (1975, p. 24), however, recommend in situations calling for non-directional alternatives (e.g., in two-group studies) that the "one-sided" P-value be reported along with the direction of the observed departure of the null hypothesis.

Many data analysis programs for use with mainframe computers and some for microcomputers yield P-values as output. Often, however, researchers rely on probability tables in the back of textbooks to determine P-values. Often, too, these tables are inadequate to give anything but very rough probability estimates. If more than a few centiles are available, the degrees-of-freedom values are often quite restricted. The same problem is encountered by those researchers wanting to determine critical values for regions of rejection. Some pretty decent tables are available for Student t and chi-squared test statistics--decent in the sense of having a fair number of centiles along with a fair number of degrees-of-freedom values. For example, Diem and Lentner (1970) give t and chi-squared tables with df = 1(1)200, and a fairly large number of centiles; for "in-between" centiles,

linear interpolation should suffice. Federighi (1959) gives t centiles out to 99.9999 for a fair number of df-values.

Typically, F tables are more of a problem. Fairly exact P-values and critical values for F distributions may, however, be determined using a relationship given by Paulson (1942). He states that an F(m , n) variable is approximately related to a standard normal variable (U) in the following way:

U = ((1-2/9n) G - (1-2/9m))/

pos sq rt ((2/9n) G**2 + 2/9m),

where G is the cube root of F. Thus, for a particular F-value, one can find the corresponding U-value and then obtain the P-value from a standard normal probability table. (See Lackritz, 1984, for another method for finding P-values.) This U-F relationship was used earlier in step F5, and later in step C7.

To determine a critical F-value for Hypothesis Testing, one needs to solve a quadratic equation in G. The equation is aG**2 + bG + c = 0, where

a = 1 - 4/9n + 4/81n**2 - 2S/9n,

b = -2 + 4/9m + 4/9n - 8/81mn, and

c = 1 - 4/9m + 4/81m**2 - 2S/9m,

with S being the square of the corresponding centile in the standard normal distribution. The cube of the appropriate root of this equation will be the approximate centile of the F distribution. (An interpolation approach could also be used; see Laubscher, 1965, and Odeh & Fox, 1975, p. 57.) For upper centiles, the U-F relationship does pretty well for n > 2. This relationship was used earlier in step NP4. Approximations may be

obtained for chi-squared centiles using a method given by Hoaglin (1977) and for t centiles using a method given by Koehler (1983).

The U-F relationship would also be helpful in t-test situations, using m = 1. For example, for those of us who favor using the Bonferroni-Dunn t approach in multiple-comparison testing, fairly exact P-values may be obtained using this relationship. The typical ".01-.05 Dunn tables" are just not enough; the extensive tables given by Bailey (1977) consider, also, only the .05 and .01 levels. It is interesting to note that O. J. Dunn herself did not include the .01-.05 tables in her coauthored text (Dunn & Clark, 1974), but rather included a fairly detailed t table.

It seems reasonable to report P-values to at least three decimal places, if at all possible. Using a statement such as "P < .05" when the P-value is actually, say .028, doesn't seem to be appropriate. When combining results of independent studies using Fisher's method of adding logarithms of P-values or Edgington's method of adding P-values (see Rosenthal, 1978), three decimal places seem to be the minimum needed.

A Hybrid Approach

Some textbook authors (e.g., Anderson & Sclove, 1978, pp. 413-414) recommend the reporting of a P-value as an interpretive agent in Hypothesis Testing. Only two textbooks were found which include the finding of a P-value as a formal step in Hypothesis Testing (Glass & Hopkins, 1984, p. 205; Iman & Conover, 1983, p. 218). The proposal to be currently offered suggests a different kind of hybrid of the notions of Hypothesis Testing and Significance Testing.

From an Hypothesis Testing viewpoint, it is desirable to have both alpha = P(Type I error) and beta = P(Type II error) "fairly low." Since alpha and beta are inversely related in that as one goes up the other goes down (other things held constant; see Brewer, in press), it would be well to have some idea as to the approximate power (i.e., 1 - beta) one has for some "important" or "interesting" alternative hypothesis characterizations, given a particular alpha. In group-comparison and correlation studies there are at least three ways of characterizing an important alternative hypothesis. In the simple two-group situation (when one might be using a t test), the popular way is to specify an important "standardized absolute mean difference." In the K-group situation, this extends to specifying a function of a noncentrality parameter (see, e.g., Cohen, 1977, p. 275; Glass & Hopkins, 1984, p. 353). A second way is to specify an important "proportion of explained variation," or an important "magnitude of strength of association"--in each case, the relationship considered is that between the grouping variable and the outcome variable. In the simple two-group situation, this may involve specifying a squared point-biserial correlation coefficient or a phi coefficient. In the K-group situation, this extends to specifying an eta-squared value or a value of the Light-Margolin (1971) index. In testing H0: rho = 0, a rho squared value would be specified. In group-comparison studies there is a third way of characterizing alternative hypotheses that deserves some serious consideration. This is what Cohen (1977, p. 21) refers to as "percent nonoverlap." It may also be viewed as a "proportion of correct classification" when predicting group membership on the

basis of outcome variable scores--Huberty and Holmes (1983) discuss this in a two-criterion-group context. The two latter ways of characterization seem preferable to that of the first way.

The current proposal is simply to use Hypothesis Testing notions (along with some researcher thought) in designing the study and, then, to use Significance Testing as a basis of making statistical inferences. Specifically, by specifying alpha, an important alternative hypothesis characterization, and beta (or power value) for this characterization, one can arrive at an approximate "optimal" number of experimental units to employ. Tables and charts of such optimal sample sizes are available for a number of statistical hypotheses (see Cohen, 1977; Odeh & Fox, 1975; Olejnik, 1984). Some sample-size tables are also available for multivariate tests (Kres, 1983).

Procedurally speaking, the proposal is as follows (with a one-way layout example):

		Example
C1.	State statistical hypotheses	H0: 4 pop. means equal H1: not H0
C2.	Select appropriate test statistic	MSH/MSE
C3.	Specify alpha and beta willing to live with	alpha = .05; beta = .30
C4.	Specify alternative hypothesis characterization of concern	eta squared = .10
C5.	Determine approximate number of experimental units needed	About 20 units per group (Cohen, 1977, p. 384)

C6.	Collect data and calculate value of test statistic	F = 4.27
C7.	Determine estimated probability (P) of obtaining such a test statistic value or one more extreme, under HO	P = P(F(3,76) > 4.27 | HO) = .008
C8.	Determine value of explained variation index	eta squared = .144
C9.	Conclude existence of "real" effects (as indicated by hypotheses) if P-value is "small" and explained variation is "substantial"	Grouping variable has an effect on outcome variable

The definitions of "small" and "substantial" in step C9 are, of course, arbitrary. Assessing the size of a P-value often follows conventions used in assessing the size of an alpha-value. (See the earlier discussion.) Assessment of a proportion of explained variation is equally arbitrary, although some recommendations have been given by Cohen (1977)--see, also, Feldt (1973).

An approach taken by some researchers may be to first decide whether or not to reject HO on the basis of the size of the P-value, and then use the eta-squared value in making an "interpretation." As indicated above, I prefer to consider the two values jointly to draw a conclusion from the results. It may

be noted that one minus eta-squared can itself be considered a test statistic for the null hypothesis as stated in step C1--in group comparison problems the hypotheses could be stated in terms of a relationship between the grouping variable and the outcome variable. The referent distribution for the complement of eta-squared is the two-parameter beta distribution (see Cureton, 1964; Hogg & Craig, 1978, p. 147).

The suggested inferential approach overcomes, at least in part, one criticism of Significance Testing; viz., the lack of concern for the probability of rejecting HO when it is false. Incidentally, this same criticism may be made of the practice of Hypothesis Testing.

If complicated designs are to be employed, some educated guesses will have to be made to complete step C5. In addition to designs calling for basic "parametric" analyses, Cohen (1977) discusses factorial designs and some designs calling for "nonparametric" analyses, as well as multiple regression. In any case, common sense and reasonableness are the basic virtues needed to employ the proposed strategy.

In describing a study it is recommended that the process of determining the sample size (specifically, steps C2-C5), along with a reference for the probability table used, be explicitly indicated. In reporting the results, one should give the value of the statistic used, the P-value, the value of the explained variation index, and, of course, the conclusion reached--steps C6-C9. Having implied the statistical hypotheses (C1) via a stated research question or a purpose statement, plus having described the variables of interest and measures thereof and the

sampling plan employed, all nine steps will have been essentially covered. In describing a study and reporting results, it should go without saying that the population of units to which inferences are to be made should be specified. In practice we often do not have a random sample, but we can judge our sample (of measures) to be like a random sample from, say, a normal distribution of measures. We can envision a population of random samples from a normal distribution that is formed in some "random" manner. We then "form the opinion that our set of data looks like a typical member of such a population of random samples" (Kempthorne, 1972, p. 185). Such an opinion can be based on a thorough description of the sample units. (Incidentally, random sampling of units does not necessarily assure independence of response measures!)

In reviewing a number of textbooks for potential classroom use and for the purpose of locating alternative presentations of concepts, some questionable coverage of terminology related to statistical testing has been found. This is turned to next.

On Some Statistical Testing Terminology

Issues are taken with a few statements and discussions regarding statistical testing often found in introductory textbooks that are, in my opinion, incorrect or at best misleading. Specific sources are not cited for fear of an implied condemnation of each book in toto. Many of the textbooks in which questionable statements are given are otherwise fine books--"a few poor statements an inferior book doesn't make." Sources of the statements, however, are available from the current author.

Hypothesis. The definition of a statistical hypothesis given in many textbooks is much too restrictive. Typical definitions

state that an hypothesis is a conjectural statement about: "the relation between two or more variables"; or "an unknown parameter"; or "one or more parameters of a population distribution." Such statements are correct, obviously, for some hypotheses or for some classes of hypotheses. But they do not, in fact, define an hypothesis in any general way and, therefore, hardly pass as definitions per se. Granted, such statements are typically used in an introductory discussion of statistical testing. But a student would be hard put to relate such definitions to some hypotheses that he or she might encounter in later study--even within the same textbook. If it is desirable to present a formal definition of a statistical hypothesis, a simply stated and appropriate one is that given by Hogg and Tanis (1977, p. 243): "an assertion about a distribution of one or more (random) variables" (parentheses added) (see also, Guenther, 1973, p. 163; Hodges & Lehmann, 1970, p. 309). This definition covers hypotheses studied in introductory univariate analyses, in multivariate analyses, and in analyses involving complete distributions.

Region of Rejection. A region of rejection is a subset of the set of all possible values of the test statistic. Sometimes--see the next subsection--for a nondirectional alternative hypothesis, a subset consists of two portions. Contrary to what is illustrated in some introductory textbooks, a region of rejection is not an area under a curve that represents a probability distribution. Of course, if the proposed strategy of statistical testing is accepted, the need for this term is eliminated.

Tails of a Test. Consider the following sequence of statements: (i) "a statistical test for which the region of rejection is in either the upper or lower tail of the sampling distribution is called a one-tailed test"; and (ii) "a one-tailed test is used whenever the experimenter makes a directional prediction, i.e., the alternative hypothesis is directional." As is well known, a definition is an "if-and-only-if" statement. As such the above sequence simply does not lead to a correct understanding. To point out, is the F test used in simple analysis of variance a one-tailed test? Is the chi-squared test using Pearson's statistic a one-tailed test? Beyond the simple one- and two-group univariate case, the notion as described above is near nonsensical. Why not drop it?

Contingency Table. It is somewhat typical to see and hear the expression "contingency table" when authors and students are referring to a rectangular array of counts or frequencies. And when possessing such an array, researchers often want to "run a chi-square." The statistic to which they are referring is, of course, Pearson's "O-E" chi-squared statistic. This statistic is appropriate when one is analyzing categorical variables, technically only when the categories are unordered--a point not often made explicit by textbook writers. A more bothersome point seldom made explicit is that Pearson's statistic can be utilized in two different research situations (see Kirk, 1984, pp. 379-387). (The goodness-of-fit situation is not being considered here.) One might be termed a "one-sample bivariate" situation. Here a single sample is observed and "scores" on two categorical variables are recorded, Interest is on assessing the dependence

of the two variables--the term "contingency table" may be appropriately applied here. The second situation might be termed a "K-sample test of proportions." This is a multiple-sample, single-outcome-variable situation. Interest is on assessing the equality of distributions of proportions. The appropriateness of the term "contingency table" in this latter situation is questionable.

These two situations are different in terms of design and in terms of analysis. As stated above, in the first situation one sample is selected, while in the second there are multiple samples. Also, in the first the totals in each margin of the data table summary are not fixed, while in the second the totals in one margin are considered fixed. In the first situation, the calculation of Pearson's statistic and associated P-value would be followed up with the calculation of an index of association such as that credited to H. Cramer. The follow-up calculations for the second situation, on the other hand, would be an eta-squared-like index (Light & Margolin, 1971), and, for more than two groups and/or more than two categories, some multiple comparisons. Furthermore, the basic research questions behind the two situations are different--for the first the question pertains to one of association, while for the latter it pertains to group comparison.

Discussion

Over the years there have been some heavy criticisms leveled against statistical testing (see, e.g., Carver, 1978; Meehl, 1978; Morrison & Henkel, 1970). Some of the criticisms against the use of, interpretation of, and emphasis on, statistical testing are

warranted. But there is nothing wrong with statistical tests themselves! When used as guides and indicators, as opposed to a means of arriving at definitive answers, they are okay.

It has been said that we ought to do away with statistical testing and focus on interval estimation. This may be acceptable for one- and two-sample mean-comparison situations, and maybe for some K-sample situations; but there seem to be other types of situations, some nonparametric, some multivariate, where interval estimation falls a little short.

Some very different alternatives to traditional statistical testing have been suggested, but appear to have little chance of unseating tradition. One is "relative likelihood analyses" proposed by Crane (1980). Another is an approach which has bounced around for some time and has been resurrected by Serlin and Lapsley (1985) that involves testing a composite null hypothesis in the form of an interval, as opposed to testing a simple null hypothesis. See also, Pitz (1978) and Marler (1980) for more on likelihood ratio analyses and composite hypotheses. Still another alternative is one advanced by Spielman (1974, p. 225). Spielman contends that it is unreasonable to base inferences on the rarity of the observed outcome of an experiment. Therefore, he suggests a modification of Significance Testing where the test statistic is an "ordered" distance-function over all possible experimental outcomes that reflects a comparative measure of whether or not a given outcome supports HO to a greater or lesser degree than does another. Of course, we repeatedly have the Bayesians suggest to us <u>the</u> method of statistical testing.

A strategy of statistical testing is proposed in the current paper that, if taken seriously, could well serve statistical-testing-types. The proposed hybrid of Hypothesis Testing and Significance Testing seems like a reasonable approach to use. The hybrid approach may be developed in the classroom using a sort of "backdoor" procedure. Some success (at least from an instructor viewpoint) has been achieved by introducing statistical testing via Significance Testing. Following discussions covering hypotheses and sampling (i.e., probability) distributions, the logic of Significance Testing seems to be picked up reasonably easy by students. After discussing a number of hypotheses (mostly involving single populations), the notions of Type I and Type II errors and probabilities of such, the notion of power, and the notion of explained variation are covered. This is followed with the idea of obtaining an "optimal" sample size. It is at this point that the instructor can "tie it all together."

References

American Psychological Association (1983). Publication manual (3rd ed.). Washington, DC: Author.

Anderson, T. W., & Sclove, S. L. (1978). An introduction to the statistical analysis of data. Boston: Houghton Mifflin.

Bailey, B. J. R. (1977). Tables of the Bonferroni t statistic. Journal of the American Statistical Association, 72, 469-478.

Bhattacharyya, G. K., & Johnson, R. A. (1977). Statistical concepts and methods. New York: Wiley.

Brewer, J. K. (1975). Reporting P values (letter to editor). The American Statistician, 29, 171.

Brewer, J. K. (in press). Behavioral statistics textbooks: Sources of myths and misconceptions? Journal of Educational Statistics.

Carlson, R. (1976). Discussion: The logic of tests of significance. Philosophy of Science, 43, 116-128.

Carver, R. P. (1978). The case against statistical significance testing. Harvard Educational Review, 48, 378-399.

Cohen, J. (1977). Statistical power analysis for the behavioral sciences. New York: Academic Press.

Cowles, M., & Davis, C. (1982). On the origins of the .05 level of statistical significance. American Psychologist, 37, 553-558.

Cox, D. R., & Hinkley, D. V. (1974). Theoretical statistics. London: Chapman and Hall.

Crane, J. A. (1980). Relative likelihood analyses versus significance tests. Evaluative Review, 4, 824-842.

Cureton, E. E. (1964). Z, t, F or B. The American Statistician, 18, 26+.

Diem, K., & Lentner, C. (Eds.) (1970). Scientific tables. Ardsley, NY: Geigy.

Dunn, O. J., & Clarke, V. A. (1974). Applied statistics. New York: Wiley.

Federighi, E. T. (1959). Extended tables of the percentage points of Student's t-distribution. Journal of the American Statistical Association, 54, 683-688.

Feldt, L. S. (1973). What size samples for methods/materials experiments? Journal of Educational Measurement, 10, 221-226.

Gibbons, J. D., & Pratt, J. W. (1975). P-values: Interpretation and methodology. The American Statistician, 29, 20-25.

Glass, G. V, & Hopkins, K. D. (1984). Statistical methods in education and psychology. Englewood Cliffs, NJ: Prentice-Hall.

Guenther, W. C. (1973). Concepts of statistical inference. New York: McGraw-Hill.

Guilford, J. P., & Fruchter, B. (1978). Fundamental statistics in psychology and education. New York: McGraw-Hill.

Hacking, I. (1965). Logic of statistical inference. London: Cambridge University.

Hays, W. L. (1981). Statistics. New York: Holt, Rinehart and Winston.

Hoaglin, D. C. (1977). Direct approximations for chi-squared percentage points. Journal of the American Statistical Association, 72, 508-515.

Hodges, J. L., & Lehmann, E. L. (1970). Basic concepts of probability and statistics. San Francisco: Holden-Day.

Hogg, R. V., & Craig, A. T. (1978). Introduction to mathematical statistics. New York: Macmillan.

Hogg, R. V., & Tanis, E. A. (1977). Probability and statistical inference. New York: Macmillan.

Huberty, C. J, & Holmes, S. E. (1983). Two-group comparisons and univariate classification. Educational and Psychological Measurement, 43, 15-26.

Iman, R. L., & Conover, W. J. (1983). A modern approach to statistics. New York: Wiley.

Kempthorne, O. (1972). Theories of inference and data analysis. In T. A. Bancroft (Ed.), Statistical papers in honor of G. W. Snedecor (pp. 167-191). Ames, IA: Iowa State University.

Kempthorne, O. (1976). Of what use are tests of significance and tests of hypotheses. Communications in Statistics, A5, 763-777.

Kempthorne, O., & Folks, L. (1971). Probability, statistics and data analysis. Ames, IA: Iowa State University.

Kirk, R. E. (1984). Elementary statistics. Monterey, CA: Brooks/Cole.

Koehler, K. J. (1983). A simple approximation for the percentiles of the t distribution. Technometrics, 25, 103-105.

Kres, H. (1983). Statistical tables for multivariate analysis. New York: Springer-Verlag.

Lackritz, J. R. (1984). Exact p values for F and t tests. The American Statistician, 38, 312-314.

Laubscher, N. F. (1965). Interpolation in F-tables. The American Statistician, 19(1), 28+.

Lawlis, G. F., & Peek, L. A. (1979). Testing the null hypothesis: An unstatement. Multivariate Experimental Clinical Research, 4, 133-137.

Light, R. J., & Margolin, B. H. (1971). Analysis of variance for categorical data. Journal of the American Statistical Association, 66, 534-544.

Marler, M. R. (1980). Likelihood ratio tests of hypotheses: Comments on Pitz's article and some alternative procedures. Psychological Bulletin, 87, 568-574.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. Journal of Consulting and Clinical Psychology, 46, 806-834.

Morrison, D. E., & Henkel, R. E. (Eds.). (1970). The significance test controversy. Chicago: Aldine.

Mosteller, F., Fienberg, S. E., & Rourke, R. E. K. (1983). Beginning statistics with data analysis. Reading, MA: Addison-Wesley.

Nagel, S. S., & Neef, M. (1977). Determining an optimal level of statistical significance. In M. Guttentag (Ed.), Evaluation studies review annual (Vol. 2) (pp. 146-158). Beverly Hills, CA: Sage.

Neyman, J. (1976). Tests of statistical hypotheses and their use in studies of natural phenomena. Communications in Statistics, A5, 737-751.

Odeh, R. E., & Fox, M. (1975). Sample size choice. New York: Dekker.

Olejnik, S. F. (1984). Planning educational research: Determining the necessary sample size. Journal of Experimental Education, 53, 40-48.

Paulson, E. (1942). An approximate normalization of the analysis of variance distribution. Annals of Mathematical Statistics, 13, 233-235.

Pitz, G. F. (1978). Hypothesis testing and the comparison of imprecise hypotheses. Psychological Bulletin, 85, 794-809.

Rosenthal, R. (1978). Combining results of independent studies. Psychological Bulletin, 85, 185-193.

Savage, L. J. (1976). On rereading R. A. Fisher. Annals of Statistics, 4, 441-500 (with discussion).

Sawyer, A. G., & Peter, J. P. (1983). The significance of statistical significance tests in marketing research. Journal of Marketing Research, 20, 122-133.

Seidenfeld, T. (1979). Philosophical problems of statistical inference. Boston: Reidel.

Serlin, R. C., & Lapsley, D. K. (1985). Rationality in psychological research: The good-enough principle. American Psychologist, 40, 73-83.

Snedecor, G. W., & Cochran, W. G. (1967). Statistical methods. Ames, IA: Iowa State University.

Spielman, S. (1974). The logic of tests of significance. Philosophy of Science, 41, 211-226.

Spielman, S. (1978). Discussion: Statistical dogma and the logic of significance testing. Philosophy of Science, 45, 120-135.

Trawinski, I. M. (1975). (Letter to editor). The American Statistician, 29, 171-172.